Revista da Estatística UFOP, Vol III(3), 2014, ISSN 2237-8111
Edição especial: 59ª Reunião Anual da Regional Brasileira da Sociedade Internacional de Biometria - RBRAS
Trabalhos apresentados na versão poster.

Modelos volumétricos de árvores com variáveis categóricas: Uma avaliação preditiva

Vinícius Ribeiro Florêncio [1]
Rogério Cecere Vaz [2]
Edgar de Souza Vismara [3]

1 Introdução

O volume é uma das informações mais importantes para determinar o potencial de um povoamento florestal já que este fornece meios de calcular o estoque de madeira e o potencial de produção da floresta (THOMAS et al. 2006).

Segundo Campos e Leite (2006), empregar modelos que determinam o volume representa uma das principais formas de se quantificar a produção e estoque de povoamentos florestais. O volume das árvores cubadas é correlacionado com variáveis como altura total e DAP. Existem vários modelos que expressam o volume se utilizando dessas variáveis, porém, o modelo de Schumacher e Hall (1933) resulta em estimativas precisas e livres de tendências.

Apesar do modelo de Schumacher-Hall possuir as características citadas acima, este pode gerar um certo viés quando os dados de cubagem possuem uma estrutura hierárquica (ZANDONÁ et al. 2008). Devido a esse problema a inserção de uma variável adicional pode ser um meio de reduzir o erro de predição.

Vale ressaltar que um problema comum no ajuste de modelos a partir de amostra envolve estimar um parâmetro θ desconhecido e apesar dos modelos gerarem estimadores aproximados não é possível determinar o quão próximo dos parâmetros esses estimadores estão. Um dos métodos para demonstrar essa aproximação é a validação cruzada (EFRON e TIBSHIRANI 1986). Muitos métodos dentro da validação cruzada foram propostos e examinados e para cada um desses métodos busca-se o valor mais próximo possível do verdadeiro erro gerado pelos modelos (KEARNS e RON 1999). Rogers e Wagner (1978) e Devroye e Wagner (1979) provaram que para vários modelos específicos o leave-one-out consegue ser um estimador que mais se aproxima do erro.

Outro método que pode trazer uma perspectiva preditiva de modelos é o critério de informação de Akaike (AIC), já que, Stone (1974) e Akaike (1974), através de uma abordagem teórica e Davies et al. (2005), através de uma abordagem empírica, demonstraram a equivalência assintótica do AIC e da validação cruzada no processo de seleção de modelos. Isso acontece pois o AIC

[1]COENF - UTFPR/DV. e-mail: *vinicius5008@gmail.com*
[2]COENF - UTFPR/DV. e-mail: *cvazrogerio@gmail.com*
[3]COENF - UTFPR/DV. e-mail: *desouzavismara@gmail.com*

é um critério de seleção que privilegiará famílias de aproximação parcimoniosa mais próximas ao modelo operacional (BUCKLAND et al. 1997).

Dentro desse contexto o objetivo desse trabalho é aumentar a precisão da predição do volume total das árvores de *Eucalyptus grandis*, incluindo uma variável categórica que representa o nível hierárquico em que as árvores se encontram. Além disso, buscou-se verificar a relação de equivalência entre o AIC e a validação cruzada.

2 Material e Métodos

O estudo foi realizado em florestas de *Eucalyptus grandis* localizadas em três diferentes cidades do estado de São Paulo: Bofete, Itatinga e Salto. Os dados resultaram num total de 7881 árvores com valores medidos de altura total, DAP e volume total.

Para obtenção dos modelos de predição do volume total das árvores partiu-se do modelo teórico de Schumacher e Hall (1933): $\ln(vt) = \beta_0 + \beta_1 \ln(ht) + \beta_2 \ln(d)$, onde vt é o volume total em (dm^3), ht é a altura total em (m) e d é o DAP em (cm). A este modelo foi incluído variáveis categóricas que representam a estrutura hierárquica do conjunto de dados. Foram elas: *região, fazenda, projeto, estrato e talhão*.

Os modelos ajustados por mínimos quadrados ordinários, foram comparados quanto ao ajuste através dos valores do Coeficiente de determinação ajustado (R^2), Erro Padrão Residual (EPR), do Critério de Informação de Akaike (AIC) e do Critério de Informação de Akaike corrigido $(AICc)$.

Ao melhor modelo selecionado, quanto ao ajuste, e ao modelo de Schumacher-Hall na sua forma tradicional(sem variável categórica) foi aplicado um procedimento de validação cruzada do tipo ”leave-one-out”(KEARNS e RON, 1999) afim de avalia-los quanto ao seu poder preditivo. Esse procedimento consiste na retirada de uma árvore por vez do conjunto de dados para realização do procedimento de ajuste e para imediatamente após, predizer a árvore retirada. Este procedimento gerou 7881 erros de predição, que posteriormente foram usados no calculo da raiz quadrada do erro médio quadrático (RMSE) do modelo.

3 Resultados e discussões

A análise comparativa entre o modelo de Schumacher-Hall e os demais modelos quanto aos valores de EPR e AICc (Tabela 1) demonstram ganho de ajuste com a entrada das variáveis categóricas. Este ganho ocorre em diferentes graus de intensidade sendo mais relevante no sítio estrato. Os valores calculados de R^2 para todos os modelos com variável categórica, à exceção do sítio *estrato*, foram inferiores ao modelo de Schumacher-Hall. Segundo Chatterjee e Hadi (2006) o R^2 traz apenas a ideia de quanto as variáveis preditoras conseguem explicar a variação da variável resposta, avaliando o modelo em relação ao ajuste e não quanto ao poder preditivo.

Ficou evidente que o sítio *estrato* apresentou os melhores resultados quanto a todos critérios de avaliação. Além disso, seus valores de EPR, AIC e AICc se destacam dessa forma escolhendo-o para aplicação da validação cruzada.

Modelo	Estimativas					
	EPR	Ganho EPR %	R^2	AIC	AICc	ΔAICc
Estrato	0.1722	16.16%	0.9778	-1370.257	-1361.39	-380.9179
Fazenda	0.1912	6.91%	0.9717	-989.9907	-989.4721	-51.53
Projeto	0.1933	5.89%	0.9711	-938.6536	-937.9421	-136.399
Talhão	0.1979	3.65%	0.9703	-805.1587	-801.5413	-1.8504
Região	0.2006	2.34%	0.9686	-799.7431	-799.6909	-99.2939
Schumacher	0.2054	-	0.9761	-700.4156	-700.397	-

Tabela 1: Dados obtidos pelos modelos ajustados, onde "Ganho EPR" é a diferença entre o modelo de Schumacher e o modelo com variável categórica em porcentagem e "ΔAICc" é a diferença entre os valores das colunas de "AICc" .

A partir do método "leave-one-out" foi possível calcular os erros de predição de cada árvore e em seguida calcular seus respectivos valores de RMSE apresentados na Tabela-2.

RMSEA	RMSEB	Ganho RMSE %
0.2499226	0.2008331	19.64%

Tabela 2: Dados obtidos pela validação cruzada, onde "RMSEA" é a raiz quadrada do erro quadrático médio de predição do modelo de Schumacher-Hall e "RMSEB" é a mesma estatística para o modelo com a adição da variável categórica *estrato*.

Esses resultados indicam o bom poder de predição que a variável categórica *estrato* trouxe para o modelo. Esse poder de predição é também evidenciado pela análise gráfica. No gráfico A da Figura-1 os erros de predição calculados para o modelo de Schumacher-Hall apresentam maior tendência de superestimativas para algumas árvores. Já o modelo que incluiu a variável categórica *estrato* (gráfico B da Figura-2), apesar de apresentar uma distribuição mais dispersa dos erros, tem um comportamento enviesado menos importante. O fato do RMSE penalizar erros maiores com maior peso foi determinante para o ganho preditivo da abordagem que incluiu a variável categórica.

É preciso ressaltar que o valor de AIC já apontava para o ganho preditivo da inclusão da variável categórica que foi confirmado pela validação cruzada. Isso confirma a equivalência entre este critério e o RMSE obtido pela validação e demonstra a capacidade deste em avaliar os modelos quanto ao poder preditivo. O fato do ganho em EPR com a inclusão da variável categórica ser maior na avaliação preditiva que na avaliação quanto ajuste ($16.16\% vs 19.64\%$) é outra razão para o AIC ser considerado um critério mais adequado para as escolha de modelos no contexto preditivo.

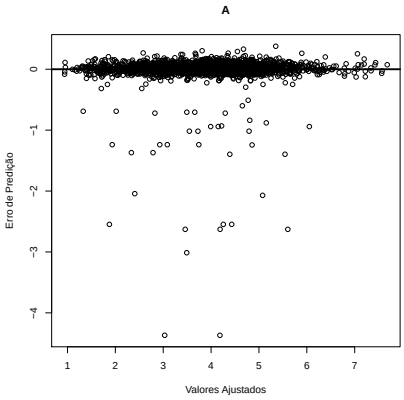

Figura 1: Distribuição dos erros de predição com validação cruzada do modelo de Schumacher-Hall.

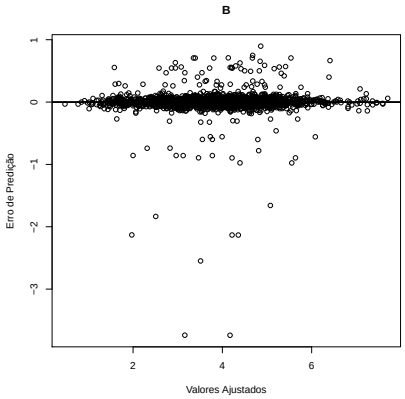

Figura 2: Distribuição dos erros de predição com validação cruzada do modelo com inclusão da variável categórica *estrato*.

4 Conclusão

O ajuste do modelo com variável categórica incrementou consideravelmente a análise do volume das observações de *Eucalyptus grandis* em relação ao modelo de Schumacher-Hall sobretudo o sítio *estrato*, onde para este sítio o AICc apresentou um valor de -1370.2516 e o EPR um ganho de 16.16%. A validação cruzada leave-one-out mostrou um resultado de predição muito significativo para o sítio estrato de 19.64% confirmando os valores anteriores.

O critério de informação de Akaike(AIC) demonstrou um alto poder em avaliar os modelo no contexto preditivo. É devido a essa característica de avaliação de modelos que este critério deve ser utilizado principalmente quando não existe a possibilidade de realizar a validação cruzada.

Os resultados mostraram a relação de equivalência entre o AICc e a validação cruzada, além da importância da inclusão de uma variável categórica para observações que apresentam uma estrutura hierárquica.

Referências

[1] AKAIKE, H. A new look at statistical model identification. **IEEE Transactions on Automatic Control**, Tokyo, v.19, n.6, p.717-723, 1974

[2] BUCKLAND, S. T.; BURNHAM, K. P.; AUGUSTIN, N. H. Model slection: A integral part of inference. **Biometrics**, London, v.53, n.2, p.603-618, 1997

[3] CAMPOS, J. C. C.; LEITE, H. G. **Mensuração florestal: perguntas e respostas**. 2.ed. Viçosa, MG: Universidade Federal de Viçosa, 2006. 470p.

[4] CHATTERJEE, S.; HADI, A. S. **Rgression analysis by example**. 4th.ed. New York: Wiley, 2006. 42p.

[5] DAVIES, S. L.; NEATH, A. A.; CAVANAUGH, J. E. Cross validation model selection criteria for linear regression based on the Kullbak-Leibler discrepancy. **Statistical Methodology**, Amsterdam, v.2, n.4, p.249-266, 2005.

[6] DEVROYE, L. P.; WAGNER, T. J. Distribution-free inequalities for the deleted and holdout error estimates. **IEEE Transactions on Information Theory**, v.25, n.2, p.202-207, 1979.

[7] EFRON, B.; TIBSHIRANI, R. Bootstrap method for standard errors, confidence intervals, and others measures of statistical accuracy. **Statiscal Science**, v.1, n.1, p.54-77, 1986.

[8] KEARNS, M.; RON, D. Algorithmic Stability and Sanity-Check Bounds for Leave-One-Out-Cross-Validation, **Neural Computation**, v.11, n.6, p.1427-1453, 1999.

[9] ROGERS, W. H.; WAGNER T. J. A fnite sample distribution-free performance bound for local discrimination rules. **The Annals of Statistics**, v.6, n.3, p.606-514, 1978.

[10] SCHUMACHER, F. X.; HALL, F. S. Logarithmic expression of timber-tree volume. **Journal of Agricultural Research**, v.47, n.9, p.719-734, 1933.

[11] STONE, M. Cross-validatory choice and assessment of statistical predoictions. **Journal of the Royal Statistical Society**. Series B(Methodological), London, v.36, n.2, p.111-147, 1974.

[12] THOMAS, C. et al. Comparação de Equações Volumétricas Ajustadas com Dados de Cubagem e Análise de Tronco, **Ciência Florestal**, Vol. 16, No. 3, p.319-327, 2006.

[13] ZANDONÁ, D. F.; LINGNAU, C.; NAKAJIMA, N. Y. Varredura a Laser aerotransportado para estimativa de variáveis dendrométricas. **Scientia Forestalis**, Piracicaba, v. 36, n. 80, p. 295-306, 2008.